EPMA Journal (2010) 1:377–387
DOI 10.1007/s13167-010-0040-3

REVIEW ARTICLE

Moving from correlative science to predictive oncology

Richard Simon

Received: 8 April 2010 / Accepted: 29 June 2010 / Published online: 22 July 2010

Abstract Many diagnostic entities traditionally viewed as individual diseases are heterogeneous in molecular pathogenesis and treatment responsiveness. This results in treatment of many patients with ineffective drugs, the conduct of large clinical trials to identify small average treatment benefits for heterogeneous groups of patients. In oncology, genomic technologies provide powerful tools for identification of patients who require systemic treatment and for selecting the most appropriate drug. Development of drugs with companion diagnostics, however, increases the complexity of clinical development and requires new approaches to the design and analysis of clinical trials. Adapting to the fundamental importance of tumor genomics will require paradigm changes for clinical and statistical investigators in academia, industry and government. In this paper we attempt to address some of these issues and to comment specifically on the design of clinical studies for evaluating the clinical utility and robustness of prognostic and predictive biomarkers.

Keywords Biomarker · Predictive · Prognostic · Clinical trial design · Validation

Introduction

The dominant themes in oncology therapeutics today are the molecular heterogeneity of different tumors of the same primary site, the development of drugs molecularly targeted to de-regulated signaling pathways, and the personalization of treatment planning. The development of molecularly targeted drugs has accelerated the movement to personalized therapeutics based on genomic characterization of individual tumors. This is particularly true in breast cancer where treatment selection is often based on estrogen receptor status, HER2 amplification status, and gene expression profile indicating the prognostic aggressiveness of the disease.

Traditionally, the term "biomarker" referred to a measurement that tracks the pace of a disease; increasing as the disease progresses and decreasing as it regresses. Such biomarkers are sometimes referred to as "surrogate" endpoints, implying that they are surrogates for survival or other measures of clinical outcome. However few disease endpoint biomarkers in oncology have been demonstrated to be more than just correlates of survival. The distinction between a correlate and a surrogate is causality. For example, tumor shrinkage after a standard treatment may be correlated with survival because patients with smaller tumors have better response rates and longer survival. Increasing response rate, however, may not result in extended survival. It is very difficult to establish that an intermediate endpoint is a true surrogate of clinical outcome [1–4]. Nevertheless, intermediate endpoint biomarkers can be useful for early clinical development of a drug without being established as valid surrogates of clinical outcome. Pharmacodynamic biomarkers are used in for establishing that the drug inhibits its intended target and intermediate endpoint biomarkers such as KI67 or PSA can be used in phase II studies for dose selection, predictive biomarker development, and determination of whether to conduct a phase III clinical trial. Such endpoints are often not, however, acceptable as endpoints for phase III clinical trials, at least not phase III registration trials.

R. Simon (✉)
National Cancer Institute,
9000 Rockville Pike,
Bethesda MD 20892-7434, USA
e-mail: rsimon@nih.gov

Our focus here will be on baseline biomarkers, not endpoint biomarkers. Prognostic markers are baseline (pre-treatment) measurements that provide information about the patient's likely long-term outcome either untreated or with standard treatment. Prognostic markers can be used to determine whether the patient requires any systematic treatment or any beyond the standard treatment. Predictive markers are baseline measurements that indicate whether the patient is likely (or unlikely) to benefit from a specific drug or regimen.

Technologies such as array based hybridization assays and next generation DNA sequencing provide molecular characterizations of individual tumors which have the potential to improve therapeutic decision making. Development of prognostic and predictive biomarkers based on this information also has great potential value for cancer drug development and for controlling medical costs by reducing the treatment of cancer patients with regimens that do not benefit them. Nevertheless, the translation of molecular profiling data into meaningful molecular targets and effective biomarkers is not straightforward. Co-development of new drugs with companion diagnostics increases the complexity of development and may not generally provide a quicker and cheaper approach as sometimes claimed. Diagnostics which are not reliably evaluated can detract from proper patient management and increase the cost of medical care. One of the greatest challenges today is to develop and evaluate prognostic and predictive biomarkers in a reliable but practical manner that permits the translation of the genomic information read from individual tumors into therapeutic strategies that benefit patients.

We will use the term "biomarker" to include both single and composite biological measurements. A single measurement may be a protein level, a transcript abundance level, a binary indicator of the presence or absence of a gene mutation, e.t.c. Composite measurements combine the values of multiple measurements into either a quantitative score or a categorical classifier. The most common kinds of composite classifiers today are based on expression levels of multiple genes like the OncotypeDx recurrence score [5]. A composite biomarker score is characterized by its components and the way the components are combined into a single score. In many cases the score is a linear combination of the components and in that case the weights assigned to the components must be specified for the score to be well defined and useable in a prospective manner. Composite biomarker scores may be transformed to composite biomarker classifiers by introducing one or more cut-points. For example, OncotypeDx is sometimes used as a classifier for low-risk, intermediate-risk and high-risk of recurrence for patients with ER positive node negative breast cancer receiving anti-estrogens as the only systemic treatment after local therapy [6]. Currently, most composite biomarkers are based on gene expression because hybridizing fluorescently labeled transcripts to solid surface DNA microarrays enabled genome-wide evaluation of transcript abundance. Many of the issues described below concerning the development and evaluation of single and composite biomarkers also apply to markers based on mutation, genome copy number variation, methylation, single protein measurements and whole genome proteomics, as well as post-translational modifications of proteins.

Prognostic biomarkers

There is a substantial gap between the vast literature of prognostic markers and the limited use of prognostic markers in clinical practice [7]. We will explore here some of the reasons for this gap.

Developmental studies of prognostic biomarkers

The vast majority of reports of gene expression based prognostic signatures are "developmental" studies in which a signature is developed. In contrast, a "validation study", to be discussed below, utilizes a fully specified marker developed in a previous developmental study.

There are many potential problems with developmental studies of prognostic signatures based on gene expression profiling [8, 9]. Subramanian and Simon [10] recently reviewed prognostic biomarker signature development for early stage lung cancer. They found several serious flaws in the studies, but perhaps the most serious was the lack of focus on an intended use of the prognostic classifier being developed. In contrast, the OncotypeDx recurrence score identifies women with node negative ER positive breast cancer receiving Tamoxifen as their only systemic therapy who have a risk of recurrence sufficiently small that they may opt not to receive cytotoxic chemotherapy [5]. Early definition of that intended use drove the planning of the developmental studies of OncotypeDx. For a prognostic marker to be useful for therapeutic decision-making, it must be developed based on study of patients selected for that intended use. In their review of lung cancer signatures, Subramanian and Simon found that few if any studies selected patients or sized their studies based on a defined intended use.

Most prognostic marker studies are conducted using a "convenience sample" of available specimens without focus on an intended use. Unless the specimens are archived from cases on a single clinical trial, they may be quite heterogeneous with regard to clinico-pathological variables that are a part of practice standards that determine treatment, and even heterogeneous with regard to treatment.

Often the publication develops a signature that is claimed to be prognostic or more prognostic than the standard staging system. Because the patients included are so heterogeneous, however, it is often very difficult to determine whether the signature has any potential value for therapeutic decision making.

Showing that a signature separates a group of well staged stage I lung cancer patients with negative margins and no other high risk features into a subgroups with low and high risk of recurrence, if performed effectively, can establish a clinical validity of the signature, but not it's "medical utility." "Clinical validation" means that the signature correlates with clinical outcome. "Medical utility", however, means that the signature is "actionable" and can be used in therapeutic decision-making in a way that results in benefit to patients. Establishing medical utility requires that the classifier provides a therapeutic decision tool that results in patient benefit compared to practice standards based on current staging and clinical/histopathologic prognostic factors. Establishing medical utility is the objective of the validation study, but the developmental study should be designed and analyzed in a manner that provides a prognostic tool with promise of having medical utility. For the stage I lung cancer example, the validation study would probably involve randomizing stage I patients to receive or not receive adjuvant chemotherapy, with separate analysis of the patients predicted to be at low risk of recurrence and those predicted to be at higher risk of recurrence.

For developmental studies of prognostic signatures based on gene expression data, the number of genes examined is much greater than the number of cases included. Consequently, it is not appropriate to use the full set of data in the traditional manner for both developing the model and for examining the predictive ability of the model. For example, when the outcome is survival or disease-free survival, it is quite misleading to show Kaplan Meier curves based on classifying the same patients used for developing the model. Simon et al. have shown that it is almost always possible to have good apparent model fit even when none of the genes have expression values correlated with outcome [11]. This was also shown graphically by Subramanian and Simon [10]. Unfortunately, such misleading analyses are prevalent in prognostic signature studies [8, 10].

Most developmental studies should include internal validation based on either splitting the cases into training and test sets or on complete cross-validation. Both the split-sample method and cross-validation provide valid estimates of prediction power when performed correctly. The most common defect of using split-sample methods is having too few cases in the test set [11]. The most common error in using cross-validation is failure to use "complete" cross-validation in which the genes for inclusion in the classifier are re-selected for each loop of the cross-validation. Dupuy and Simon [8] found in their review that approximately 33% of studies used cross-validation incorrectly. Subramanian and Simon have developed a checklist of key features that physicians should look for in reports of prognostic gene expression signatures [9].

Split-sample validation or complete cross-validation are important components for developmental studies, but they represent internal validations that do not incorporate many components of variability that will be found in a prospective validation study or in clinical practice. Prior to conducting a validation study, an analytically validated assay should be developed. Analytical validation means that the assay measures what it is supposed to measure for cases in which a gold-standard assay exists. For other assays like gene expression classifiers, analytical validation means that the assay is reproducible and robust.

Validation studies of prognostic biomarkers

Ideally a prognostic classifier will be validated in a prospective clinical trial before it is accepted for broad clinical use [12]. The marker strategy design, shown in Fig. 1, is sometimes considered for evaluating the medical utility of a diagnostic test. With this design patients are randomized to be tested or not. For those who are not tested, their treatment is determined based on practice standards. For those patients randomized to be tested, the results of the test can be used in conjunction with standard prognostic factors to inform treatment decisions. The marker strategy design is often a poor choice of design, however. It is inefficient because many patients receive the same treatment regardless of which group they are randomized to [13–15]. In order to have reasonable statistical power to detect differences in outcome among the two randomization groups as a whole, a very large number of patients may have to be randomized [16]. For example, suppose the endpoint is survival disease-free beyond 5 years and that a proportion π of the patients receive the same treatment regardless of which arm they are randomized to. If we want to detect a difference Δ in the probability of 5-year DFS for patients receiving different treatments, then we would have to power the study to detect a difference of $(1-\pi)\Delta$ between the randomization groups. Since the required sample size is generally inversely proportional to the square of the difference to be detected between the randomized groups, the required sample size will be enormous if π is substantial. This inefficiency is particularly problematic for prognostic markers for identifying low risk patients for whom chemotherapy may be withheld because the prospective study is a therapeutic equivalence trial involving a small value of Δ. For example,

to have 90% power (with 5% one-sided significance) for detecting a 5 percentage point increase in the recurrence rate (Δ=.05) from a baseline of 10% would require a randomized trial of approximately 2,460 low risk patients if all patients are tested and low risk patients selected for randomization. The marker strategy design of Fig. 1, however, would require approximately 9,320 randomized patients to have 90% power for detecting the 2.5 percentage point increase in recurrence rate expected if only half of the patients are low risk based on the marker.

The marker strategy design may also be insufficiently informative in cases where the test is not just binary and the test based treatment strategy is complex. For example, suppose that patients with a low value of the marker have chemotherapy withheld, patients with intermediate values receive standard chemotherapy, and patients with high values receive intensified chemotherapy. Because the test is not performed for patients in the control group, one cannot examine results for the subsets of patients defined by test result. One is limited to just comparing the randomization groups overall.

The defects in the marker strategy design can be avoided by testing all patients and only randomizing patients for whom the treatment assignment is influenced by marker result. This modified marker strategy design, shown in Fig. 2, is currently being used in the MINDACT study to evaluate a 70 gene prognostic signature for determining whether to utilize chemotherapy for women with node negative estrogen receptor positive breast cancer [17].

The TAILORx study is a prospective clinical trial for evaluating the OncotypeDx gene expression recurrence score for women with node negative estrogen receptor positive breast cancer. The main objective of the trial is to determine whether women with a low recurrence score can have a low risk of disease recurrence even if chemotherapy is withheld. In the trial, such women are not randomized but just have chemotherapy withheld. If the recurrence score is accurate, the relapse rate for these patients would be very low and hence the potential benefit of chemotherapy very small in absolute terms [6]. In the MINDACT trial, women for whom the practice standard indicates chemotherapy but who have a low risk of recurrence based on the genomic signature are randomized between a chemotherapy arm and a no-chemotherapy arm. Nevertheless, the signature will be considered validated if the 5-year distant metastasis-free survival rate is greater than 92% in the women randomized to having chemotherapy withheld.

By validating these prognostic signatures in a fully prospective manner rather than by using archived tissue from a previously conducted series, one assures that an adequate number of patients are studied, that assay results are available on all patients, that the analysis is focused on a single pre-specified hypotheses, and that assay results reflect real-world tissue handling and laboratory variation. Such studies are expensive and time consuming however. In some cases effective validation of a classifier predictive of low recurrence risk can be accomplished with specimens

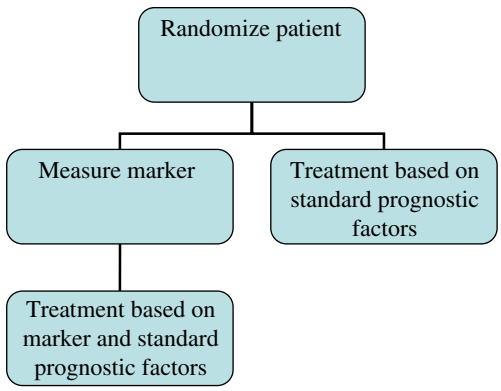

Fig. 1 The marker strategy design randomizes eligible patients between two treatment assignment strategies [46]. The control arm determines treatment using practice standards based on staging and existing prognostic factors. The new biomarker is not measured for patients randomized to the control arm. Patients randomized to the experimental arm have the candidate biomarker measured and it is used in conjunction with staging and other prognostic factors to determine treatment. This design is very flexible, but often very inefficient in the sense that the same objectives can be obtained with many fewer patients with other designs

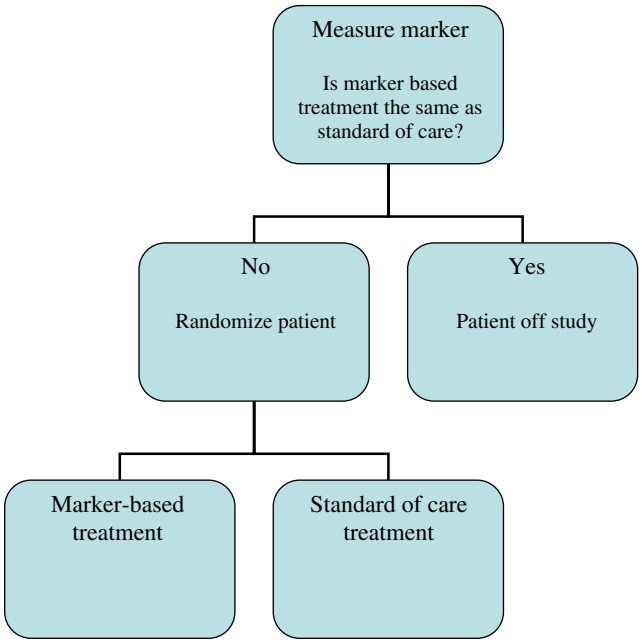

Fig. 2 The marker discordance design measures the candidate marker on all eligible patients [46]. Before randomization the practice standard determined treatment and the marker based treatment are identified. Only patients for whom the two treatments differ are randomized. This design is generally much more efficient than the marker strategy design

archived from an appropriate clinical trial that withheld chemotherapy from such patients. Convincing results are only possible, however, if the patients being analyzed are participants in a clinical trial with a design that enables unbiased evaluation of the test, if the number of patients is sufficiently large, if the proportion with available specimens adequate for testing is high, if careful analytical validation provides assurance that assay results on archived samples are accurate predictors of assay results on fresh tissue, and if the assays are blinded to clinical data [18]. When there are two or more such "prospective-retrospective studies" that satisfy these optimal conditions, Simon et al. argue that the evidence for medical utility of the prognostic biomarker should be considered commensurate with that from a fully prospective study [18]. These issues are also discussed by Pepe et al. [19].

A prognostic biomarker can also be used to identify patients whose outcome is very poor with standard chemotherapy. Although such patients may be good candidates for experimental regimens, unless there is a viable therapeutic option, such prognostic biomarkers may not be widely used in general practice.

Predictive biomarkers

Predictive biomarkers identify patients who are likely or unlikely to benefit from a specific treatment. For example, HER2 amplification is a predictive marker for benefit from trastuzumab and perhaps also from doxorubicin [20, 21] and taxanes [22]. A predictive biomarker can also be used to identify patients who are poor candidates for a particular drug; for example, advanced colorectal cancer patients whose tumors have KRAS mutations appear to be poor candidates for treatment with EGFR antibodies [23, 24].

Predictive biomarkers may be based on single gene or protein measurements, on gene expression classifiers, on pathway activation indicators, or on disease subclassifications. Measurements based on single gene or protein measurement are often closely linked to the mechanism of action of the drug. In some cases, a specific target of the drug is known but it is not clear how to best measure whether the target is driving tumor growth and invasion for an individual patient. For example, although trastuzumab was initially developed using a test for protein expression of HER2, subsequent classification has often been based on a combination of protein expression and gene amplification [25]. In the case of anti-EGFR antibodies for treatment of advanced colorectal cancer, KRAS mutation status rather than EGFR protein expression proved to be the more important predictive biomarker. In other cases, the drug has multiple molecular targets and there will be more options for determining how best to predict effectiveness of treatment. Ideally, the candidate predicted biomarkers will be evaluated during phase II clinical trials of a new drug so that a single analytically validated candidate can be used in the prospective phase III trial.

In recognition of the molecular heterogeneity of cancer, many cancer drugs are being developed today with companion diagnostics to be used as predictive biomarkers. Sawyers has stated "One of the main barriers to further progress is identifying the biological indicators, or biomarkers, of cancer that predict who will benefit from a particular targeted therapy [26]." This increases the complexity of drug development, although it has potential benefits for patients and for controlling medical expenses. It requires, however, that an effective predictive biomarker be identified and a test for it analytically validated prior to the launch of the phase III pivotal clinical trials of the drug. The discovery and phase II refinement of predictive biomarkers can be complex. It may require larger phase II databases, require new approaches to phase II trial design such as designs based on neoadjuvant treatment [7, 27, 28]

Establishing the medical utility of a companion diagnostic predictive marker for a new drug will generally be based on the phase III pivotal trials used to establish the effectiveness of the drug. In the following sections we will review some designs for phase III clinical trials that utilize new drugs and companion diagnostics.

Enrichment Designs

With an enrichment design a diagnostic test, is used to restrict eligibility for a randomized clinical trial comparing a regimen containing a new drug to a control regimen. This approach, shown in Fig. 3, was used for the development of trastuzumab in which patients with metastatic breast cancer whose tumors expressed HER2 in an immunohistochemistry test were eligible for randomization [29]. Simon and Maitournam [30–32] studied the efficiency of this approach relative to the standard approach of randomizing all patients without using the test at all. They found that the efficiency of the enrichment design depended on the prevalence of test positive patients and on the effectiveness of the new treatment in test negative patients. When fewer than half of the patients are test positive and the new treatment is relatively ineffective in test negative patients, the number of randomized patients required for an enrichment design is often dramatically smaller than the number of randomized patients required for a standard design. For example, if the treatment is completely ineffective in test negative patients, then the ratio of number of patients required for randomization in the enrichment design relative to the number

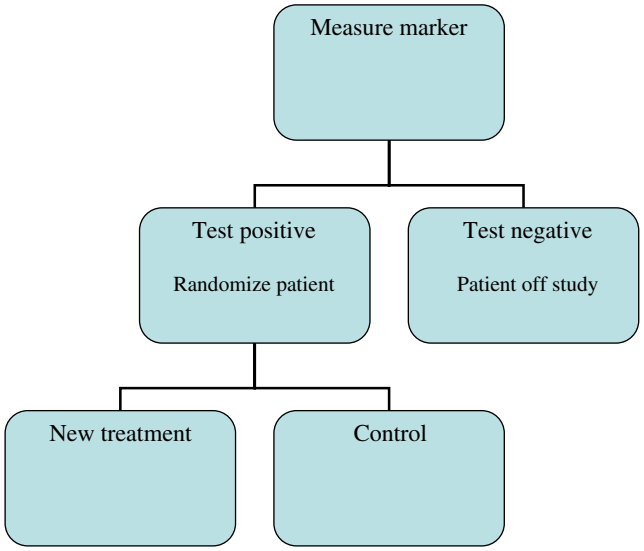

Fig. 3 The targeted enrichment design is used for evaluating a new treatment in the population of patients who are identified using a predictive biomarker as best candidates for potential benefit from the new treatment [46]. It is primarily for settings where there is a compelling basis for not expecting that "marker negative" patients can benefit from the new treatment and an analytically accurate test is available. The compelling basis is generally based on biology but could be based on substantial prior evidence with the new treatment. When the proportion of marker positive patients is less than one-half, this design can require substantially fewer randomized patients than the standard design

required for the standard design is approximately $1/\gamma^2$ where γ denotes the proportion of patients who are test positive [30]. The treatment may have some effectiveness for test negative patients either because the assay is imperfect for measuring deregulation of the putative molecular target or because the drug has anti-tumor off-target effects. Even if the new treatment is half as effective in test negative patients as in test positive patients, however, the randomization ratio is approximately $4/(\gamma+1)^2$. This equals about 2.56 when $\gamma = 0.25$, i.e. 25% of the patients are test positive, indicating that the enrichment design reduces the number of required patients to randomize by a factor of 2.56.

Hoering et al. [14] concluded that a targeted enrichment design is most efficient where there is an underlying true predictive marker and the cut point for determining the marker status is well established. Mandrekar and Sargent [13, 33] have also pointed out the efficiency of the enrichment design and suggested that the enrichment design is appropriate when (i) the new treatment has a modest absolute benefit in unselected patients but causes significant toxicity; (ii) an unselected design is ethically impossible based on previous studies; (iii) there is compelling preliminary evidence to suggest that patients without that marker profile do not benefit from the treatment; and (iv) assay reproducibility and accuracy is well established.

Zhao and Simon have made the methods of sample size planning for the design of enrichment trials available on line at http://brb.nci.nih.gov. The web-based programs are available for binary and survival/disease-free survival endpoints. The planning takes into account the performance characteristics of the tests and specificity of the treatment effects. The programs provide comparisons to standard non-enrichment designs based on the number of randomized patients required and the number of patients needed for screening to obtain the required number of randomized patients.

The enrichment design was very effective for the development of trastuzumab and the enrichment design is particularly appropriate for contexts where there is such a strong biological basis for believing that test negative patients will not benefit from the new drug that including them in would raise ethical concerns. The enrichment design does not provide data on the effectiveness of the new treatment compared to control for test negative patients. Consequently, unless there is phase II data on the clinical validity of the test for predicting response or compelling biological evidence that the new drug is not effective in test negative patients, the enrichment design may not be adequate to support approval of the test.

Designs that include both test positive and test negative patients

When a predictive classifier has been developed but there is not compelling biological or phase II data that test negative patients do not benefit from the new treatment, it is generally best to include both classifier positive and classifier negative in the phase III clinical trials comparing the new treatment to the control regimen as shown in Fig. 4. In this case it is essential that an analysis plan be

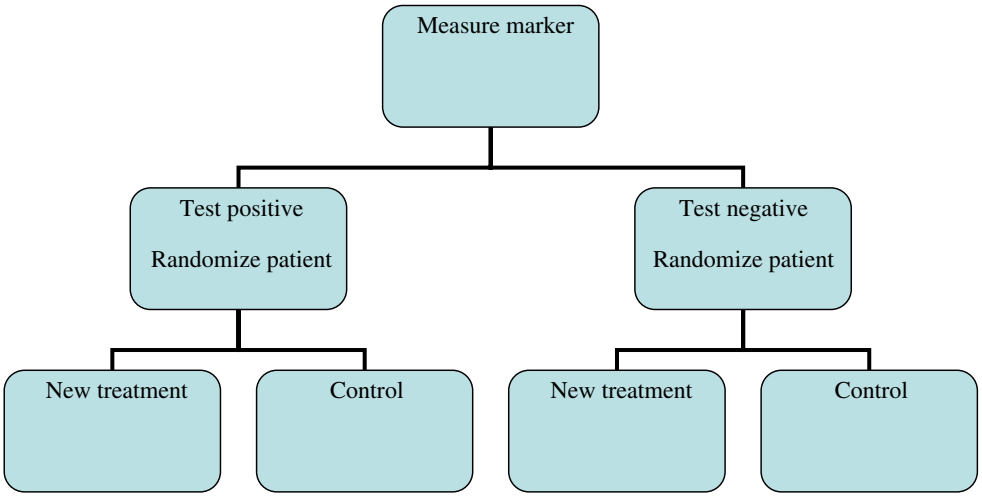

Fig. 4 The marker stratification design is used for evaluating the effectiveness of a new treatment versus a control in a population prospectively characterized by a binary predictive biomarker [46]. A detailed prospective plan should describe the primary comparison of treatment to control overall and in the marker positive and marker negative subsets. Several analysis plans are described in the text. With a focused analysis plan, claims of treatment effectiveness in marker positive patients need not be restricted to cases where the treatment is effective overall for all patients. Ideally a single completely defined analytically defined binary biomarker will be determined prior to the randomized trial. Adaptive modifications of the stratified design in which the biomarker is refined based on trial data are described in the text

pre-defined in the protocol for how the predictive classifier will be used in the analysis. It is not sufficient to just stratify, i.e. balance, the randomization with regard to the classifier without specifying a complete analysis plan. The main value of "stratifying" (i.e. balancing) the randomization is that it assures that only patients with adequate test results will enter the trial. Pre stratification of the randomization is not necessary for the validity of inferences to be made about treatment effects within the test positive or test negative subsets. If an analytically validated test is not available at the start of the trial but will be available by the time of analysis, then it may be preferable not to pre-stratify the randomization process. Similarly, if the predictive biomarker to be used in the analysis is not completely settled by the start of the trial but will be determined based on external data by the time of analysis, then careful pre-specification of the analysis plan will be necessary, but pre-stratification of the randomization process will not be appropriate.

Sargent et al. have compared the efficiency of the marker stratified design described here and shown in Fig. 4 to the marker-based strategy design shown in Fig. 1 [34]. In general, the strategy design is very inefficient for the reasons discussed above in the section on prognostic markers. Pusztai and Hess have also discussed the stratified design and the marker-based strategy [35].

The purpose of the phase III trial is to evaluate the new treatment overall and in the subsets determined by the pre-specified classifier. The purpose is not to modify or optimize the classifier. If the classifier is a composite gene expression based classifier, the purpose of the design is not to re-examine the contributions of each gene. If one does any of this, then an additional phase III trial may be needed to evaluate treatment benefit in subsets determined by the new classifier. Several primary analysis plans are presented below to illustrate that the plan should stipulate in detail how the predictive biomarker will be used in the analysis and that there should be no exploratory aspect to the treatment evaluation. These strategies are discussed in greater detail by Simon [36, 37] and a web-based tool for sample size planning with these analysis plans is available at http://brb.nci.nih.gov.

Analysis plan for biomarker with strong credentials

If one does not expect the treatment to be effective in the test negative patients unless it is effective in the test positive patients, one might first compare treatment versus control in test positive patients using a threshold of significance of 5%. Only if the treatment versus control comparison is significant at the 5% level in test positive patients, will the new treatment be compared to the control among test negative patients, again using a threshold of statistical significance of 5%. This sequential approach controls the overall type I error at 5%.

To have 90% power in the test positive patients for detecting a 50% reduction in hazard for the new treatment versus control at a two-sided 5% significance level requires about 88 events of test positive patients. If at the time of analysis the event rates in the test positive and test negative strata are about equal, then when there are 88 events in the test positive patients, there will be about $88(1-\gamma)/\gamma$ events

in the test negative patients where γ denotes the proportion of test positive patients. If 25% of the patients are test positive, then there will be approximately 264 events in test negative patients. This will provide approximately 90% power for detecting a 33% reduction in hazard at a two-sided significance level of 5%. In this case, the trial will not be delayed compared to the enrichment design, but a large number of test negative patients will be randomized, treated and followed on the study rather than excluded as for the enrichment design. This will be problematic if one does not, a-priori, expect the new treatment to be effective for test negative patients. In this case it will be important to establish an interim monitoring plan to terminate accrual of test negative patients when interim results and prior evidence of lack of effectiveness makes it no longer viable to enter them.

Fall-back analysis plan

In the situation where one has limited confidence in the predictive marker it can be effectively used for a "fall-back" analysis. Simon and Wang [38] proposed an analysis plan in which the new treatment group is first compared to the control group overall. If that difference is not significant at a reduced significance level such as 0.03, then the new treatment is compared to the control group just for test positive patients. The latter comparison uses a threshold of significance of 0.02, or whatever portion of the traditional 0.05 not used by the initial test.

If the trial is planned for having 90% power for detecting a uniform 33% reduction in overall hazard using a two-sided significance level of.03, then the overall analysis will take place when there are 297 events. If the test is positive in 25% of patients and the event rates in test positive and test negative patients are about equal at the time of analysis, then when there are 297 overall events there will be approximately 75 events among the test positive patients. If the overall test of treatment effect is not significant, then the subset test will have power.75 for detecting a 50% reduction in hazard at a two-sided.02 significance level. By delaying the treatment evaluation in the test positive patients power.80 can be achieved when there are 84 events and power.90 can be achieved when there are 109 events in the test positive subset.

Wang et al. have shown that the power of this approach can be improved by taking into account the correlation between the overall significance test and the significance test comparing treatment groups in the subset of test positive patients [39]. So if, for example a significance threshold of.03 has been used for the overall test, the significance threshold used for the subset can be somewhat greater than.02 and still have the overall chance of a false positive claim of any type limited to 5%.

Adaptive clinical trial designs using predictive biomarkers

Adaptively modifying types of patients accrued

Wang et al. [39] proposed a phase III design comparing a new treatment to a control which starts with accruing both test positive and test negative patients. An interim analysis is performed evaluating the new treatment in the test negative patients. If the observed efficacy for the control group exceeds that for the new treatment group and the difference exceeds a futility boundary, then accrual of test negative patients terminates and accrual of additional test positive patients are substituted for the un-accrued test negative patients till the originally planned total sample size is reached. Wang et al. show computer simulations that indicate this design has greater statistical power than non adaptive approaches, but their design involves many more test positive patients and may require much longer trial duration.

Liu et al. proposed a two-stage design [40] in which only marker positive patients are accrued during the initial stage. At the end of the first stage an interim analysis is performed comparing outcome for the new treatment versus control for the marker positive patients. If the results are not promising for the new treatment, then accrual stops and no treatment benefit is claimed. If the results are promising for the marker positive patients at the end of the first stage, then accrual continues for marker positive patients and accrual also commences for marker negative patients in the second stage.

Adaptive threshold design

Jiang et al. [41] reported on a "Biomarker Adaptive Threshold Design" for situations where a specific predictive index, or biomarker score, is available at the start of the trial, but a cut-point for converting the score to a binary classifier is not established. With their design, tumor specimens are collected from all patients at entry, but the value of the predictive index is not used as an eligibility criteria. Their analysis plan does not stipulate that the assay for measuring the index needs to be performed at the time of randomization. Jiang et al. [41] described two analysis plans. Analysis plan A uses the two-stage fall-back strategy described above. It begins with comparing outcomes for all patients receiving the new treatment to those for all control patients. If this difference in outcomes is significant at a pre-specified significance level α_1 then the new treatment is considered effective for the eligible population as a whole. Otherwise, a second stage test is performed using significance threshold α_2=.05–α_1. The second stage test involves finding the cut-point b* for which the treatment versus control difference in outcome (i.e. the treatment effect) is

maximized when the comparison is restricted to patients with predictive index above that cut-point. The statistical significance of that maximized treatment effect is determined by generating the null-distribution of the maximized treatment effect under random permutations of the treatment labels. If the maximized treatment effect is significant at the $1-\alpha_2$ 'th percentile of this null distribution, then the test treatment is considered effective for the subset of patients with biomarker value above the cut-point at which the maximum treatement effect occurred. This concept of using a global test to account for the multiple target populations examined can also be applied for evaluating multiple binary predictive biomarker candidates B_1, B_2, …, B_K rather than for optimizing the cut-point for a single biomarker.

Predictive analysis of clinical trials

Freidlin and Simon [42] proposed a very flexible design for a phase III trial that can be used when no classifier is available at the start of the trial. The design provides for development of the classifier and evaluation of treatment effects in a single trial while preserving the principle of separating the data used for developing a classifier from the data used for evaluating treatment in subsets determined by the classifier. It provides for using the data from a single randomized clinical trial for development and validation of a model that predicts outcome for each treatment in a randomized clinical trial based on clinical-histopathological covariates as well as biomarker covariates.

At the conclusion of the trial the new treatment is compared to the control overall using a threshold of significance of α_1 which is somewhat less than 0.05. A finding of statistical significance at that level is taken as support of a claim that the treatment is broadly effective. At that point, no biomarkers have been tested on the patients, although patients must have tumor specimens collected to be eligible for the clinical trial.

If the overall treatment effect is not significant at the α_1 level then a second stage of analysis takes place. The patients are divided into a training set and testing set. The data for patients in the training set is used to define a single subset of patients who are expected to be most likely to benefit from the new treatment compared to the control. Freidlin and Simon [42] used a machine learning algorithm based on screening thousands of genes for those with expression values that interact with treatment effect but the design can be used with other algorithms and even with candidate classifiers that do not involve gene expression. When that subset is explicitly defined, the new treatment is compared to the control for patients in the test set who have the characteristics defined by that subset. The comparison of new treatment to control for the subset is restricted to patients in the test set in order to preserve the principle of separating the data used to develop a classifier from the data used to test treatment effects in subsets defined by that classifier. The comparison of treatment to control for the subset uses a threshold of significance of $\alpha-\alpha_1$ in order to assure that the overall chance of a false positive conclusion is no greater than 0.05.

Freidlin et al. [43] have recently shown how to improve the statistical power of the approach by using k-fold cross-validation instead of simple sample splitting in the discovery and validation of the subset of patients who benefit from the new treatment. This powerful analysis strategy can be used more broadly than in the context of identifying de-novo gene expression signatures. It can be used with traditional clinico-histopathological prognostic factors or with single gene/protein candidate markers [44].

To illustrate their approach they used publicly available data for gene expression profiles and clinical outcome for 124 hormone receptor negative breast cancer patients treated on a randomized phase III neo-adjuvant clinical trial (EORTC 10994) that compared non-taxane regimen FEC (5-fluorouracil, cyclophosphamide, epirubicin) with a taxane regimen TET (epirubicin, docetaxel) [45]. The cross-validated signature design (CVASD) analysis was applied to these data and results are presented in Table 1. Although there was no overall difference in cPR rates between the TET and FEC arms (pCR rates 45% and 42%, respectively *p*-value.79). The CVASD algorithm indicated existence of a significant (*p*-value.006) subset where TET is substantially more effective than FEC: The conservative cross-validated estimate of the treatment effect was 83% pCR (TET) vs. 29% pCR (FEC). They noted that two of the probes in the signature (Hs.310359.0.A1_3p_at and g4507484_3p_a_at) are related to the MAPK pathway that has been reported to be associated with anthracycline resistance in hormone receptor negative breast cancer.

Table 1 Results of CVASD application to Bonnefoi et al. EORTC 10994 Neoadjuvant breast cancer data [45]. Analysis described in Freidlin et al. [43]

Overall comparison		
p-value 0.79		
Arm	Observed pCR rate (%) (number of patients)	
FEC	42% (66)	
TET	45% (58)	
Sensitive subset comparison		
p-value 0.006		
Arm	Estimates of pCR rates in the sensitive subpopulation	
	Re-substitution	CV
FEC	20% (15)	29% (14)
TET	100% (8)	83% (12)

Conclusions

Developments in cancer genomics and biotechnology are improving the opportunities for development of more effective therapeutics and molecular diagnostics to guide the use of those drugs. These opportunities have important potential benefits for patients and for containing healthcare costs. One of the greatest opportunities is developing predictive biomarkers of the patients who require treatment and are (or are not) likely to benefit from specific drugs.

Co-development of drugs and companion diagnostics adds complexity to the development process however. Traditional post-hoc correlative science paradigms do not provide an adequate basis for reliable predictive medicine. New paradigms are required for separating biomarker development from therapeutic evaluation. Without rigorous validation based on intended use, oncology could be inundated with expensive tests of uncertain medical utility. New clinical trial designs are required that incorporate prospective analysis plans that provide flexibility in identifying the appropriate target population in a manner that preserves overall false positive error rates. Such analysis plans must be constructed to provide information about the specificity of treatment effects without requiring such large sample sizes as to discourage development of predictive biomarkers or to require physicians to expose large numbers of patients to drugs from which they are not expected to receive benefit.

We have tried to describe effective approaches for reliable evaluation of prognostic and predictive biomarkers. These approaches include targeted enrichment designs for settings where biological evidence or phase II data destroy the equipoise necessary to include test-negative patients in the phase III clinical trial. We have emphasized that for designs that do not use the predictive biomarker as an exclusion criterion, it is essential to have a specific prospectively defined analysis plan outlining exactly how the new treatment will be evaluated with regard to the test. Because of the complexity of the biology of chronic diseases such as cancer, it is not always feasible to identify a single appropriate candidate predictive biomarker and develop an analytically validated test by the initiation of the phase III pivotal trials of a new drug. We have described several prospectively planned adaptive designs for utilizing the trial data to refine the biomarker and provide valid phase III level analyses of treatment effects.

Adapting to the fundamental heterogeneity of many human diseases and achieving the benefits of personalized predictive medicine for patients and for the economics of healthcare will require paradigm changes for academic clinical investigation, industry drug development, and for regulatory evaluation. We have attempted to identify some of the key issues involved and to provide some guidance on the design of clinical trials for evaluating the clinical utility of prognostic and predictive biomarkers.

References

1. Torri V, Simon R, Russek-Cohen E, et al. Relationship of response and survival in advanced ovarian cancer patients treated with chemotherapy. J Natl Cancer Inst. 1992;84:407.
2. Fleming TR, DeMets DL. Surrogate end points in clinical trials: are we being misled? Ann Intern Med. 1996;125:605–13.
3. Buyse M, Molensberghs G, Burzykowski T, et al. The validation of surrogate endpoints in meta-analyses of randomized experiments. Biostatistics. 2000;1:49–67.
4. Korn EL, Albert PS, McShane LM. Assessing surrogates as trial endpoints using mixed models. Stat Med. 2004;24:163–82.
5. Paik S, Shak S, Tang G, et al. A multigene assay to predict recurrence of tamoxifen-treated, node-negative breast cancer. N Engl J Med. 2004;351:2817–26.
6. Sparano JA, Paik S. Development of the 21-gene assay and its application in clinical practice and clinical trials. J Clin Oncol. 2008;26:721–8.
7. Pusztai L. Perspectives and challenges of clinical pharmacogenomics in cancer. Pharmacogenomics. 2004;5:451–4.
8. Dupuy A, Simon R. Critical review of published microarray studies for cancer outcome and guidelines on statistical analysis and reporting. J Natl Cancer Inst. 2007;99:147–57.
9. Subramanian J, Simon R (2010) What physicians should look for in evaluating prognostic gene expression signatures. Nat Rev Clin Oncol. (In Press).
10. Subramanian J, Simon R (2010) Gene expression-based prognostic signatures in lung cancer: ready for clinical use? J Nat Cancer Inst. (In Press).
11. Simon R, Radmacher MD, Dobbin K, et al. Pitfalls in the use of DNA microarray data for diagnostic and prognostic classification. J Natl Cancer Inst. 2003;95:14–8.
12. Simon R. When is a genomic classifier ready for prime time? Nat Clin Pract Oncol. 2004;1:2–3.
13. Mandrekar S, Grothey A, Goetz M, et al. Clinical trial designs for prospective validation of biomarkers. Am J Pharmacogenomics. 2005;5:317–25.
14. Hoering A, LeBlanc M, Crowley J. Randomized phase III clinical trial designs for targeted agents. Clin Cancer Res. 2008;14: 4358–67.
15. Mandrekar S, Sargent D. Clinical trial designs for predictive biomarker validation: one size does not fit all. J Biopharm Stat. 2009;19:530–42.
16. Freidlin B, McShane LM, Korn EL. Randomized clinical trials with biomarkers: design issues. J Natl Cancer Inst. 2010;102:152–60.
17. Bogaerts J, Cardoso F, Buyse M, et al. Gene signature evaluation as a prognostic tool: challenges in the design of the MINDACT trial. Nat Clin Pract Oncol. 2006;3:540–51.
18. Simon RM, Paik S, Hayes DF (2009) Use of archived specimens in evaluation of prognostic and predictive biomarkers. (Submitted for publication).
19. Pepe MS, Feng Z, Janes H, et al. Pivotal evaluation of the accuracy of a classification biomarker: the PRoBE study design. J Natl Cancer Inst. 2008;100:432–8.
20. Hayes DF. Prognostic and predictive factors revisited. Breast. 2005;14:493–9.
21. Gennari A, Sormani MP, Pronzato P, et al. HER2 status and efficacy of adjuvant anthracyclines in early breast cancer: a pooled analysis of randomized clinical trials. J Natl Cancer Inst. 2008;100:14–20.

22. Hayes DF, Thor AD, Dressler LG, et al. HER2 and response to paclitaxel in node-positive breast cancer. N Engl J Med. 2007;357:1496–506.
23. Amado RG, Wolf M, Peeters M, et al. Wild-type KRAS is required for panitumumab efficacy in patients with metastatic colorectal cancer. J Clin Oncol. 2008;26:1626–34.
24. Karapetis CS, Khambata-Ford S, Jonker DJ, et al. K-ras mutations and benefit from cetuximab in advanced colorectal cancer. N Engl J Med. 2008;359:1757–65.
25. Wolff AC, Hammond EH, Schwartz JN, et al. American Society of Clinical Oncology/College of American Pathologists Guideline Recommendations for Human Epidermal Growth Factor Receptor 2 Testing in Breast Cancer. J Clin Oncol. 2007;25: 118–45.
26. Sawyers CL. The cancer biomarker problem. Nature. 2008; 452:548–52.
27. Hess KR, Anderson K, Symmans WF, et al. Pharmacogenomic predictor of sensitivity to preoperative paclitaxel and 5-flurouracil, doxorubicin, cyclophosphamide chemotherapy in breast cancer. J Clin Oncol. 2006;24:4236–44.
28. Pusztai L, Anderson K, Hess KR. Pharmacogenomic predictor discovery in phase II clinical trials for breast cancer. Clin Cancer Res. 2007;13:6080–6.
29. Slamon DJ, Leyland-Jones B, Shak S, et al. Use of chemotherapy plus a monoclonal antibody against HER2 for metastatic breast cancer that overexpresses HER2. N Engl J Med. 2001;344: 783–92.
30. Maitournam A, Simon R. On the efficiency of targeted clinical trials. Stat Med. 2005;24:329–39.
31. Simon R, Maitournam A. Evaluating the efficiency of targeted designs for randomized clinical trials. Clin Cancer Res. 2005;10: 6759–63.
32. Simon R, Maitournam A. Evaluating the efficiency of targeted designs for randomized clinical trials: supplement and Correction. Clin Cancer Res. 2006;12:3229.
33. Mandrekar S, Sargent D (2009) Clinical trial designs for predictive biomarker validation: Theoretical considerations and practical challenges. J Clin Oncol. (In Press).
34. Sargent DJ, Conley BA, Allegra C, et al. Clinical trial designs for predictive marker validation in cancer treatment trials. J Clin Oncol. 2005;23:2020–7.
35. Pusztai L, Hess KR. Clinical trial design for microarray predictive marker discovery and assessment. Ann Oncol. 2004;15:1731–7.
36. Simon R. Using genomics in clinical trial design. Clin Cancer Res. 2008;14:5984–93.
37. Simon R. Designs and adaptive analysis plans for pivotal clinical trials of therapeutics and companion diagnostics. Expert Rev Mol Diagn. 2008;2:721–9.
38. Simon R, Wang SJ. Use of genomic signatures in therapeutics development. Pharmacogenomics J. 2006;6:1667–173.
39. Wang SJ, O'Neill RT, Hung HMJ. Approaches to evaluation of treatment effect in randomized clinical trials with genomic subset. Pharm Stat. 2007;6:227–44.
40. Liu A, Li Q, Yu KF, et al (2009) A threshold sample-enrichment approach in a clinical trial with heterogeneous subpopulations. Stat Med. (In Press).
41. Jiang W, Freidlin B, Simon R. Biomarker adaptive threshold design: a procedure for evaluating treatment with possible biomarker-defined subset effect. J Natl Cancer Inst. 2007;99: 1036–43.
42. Freidlin B, Simon R. Adaptive signature design: an adaptive clinical trial design for generating and prospectively testing a gene expression signature for sensitive patients. Clin Cancer Res. 2005; 11:7872–8.
43. Freidlin B, Jiang W, Simon R. The cross-validated adaptive signature design for predictive analysis of clinical trials. Clin Cancer Res. 2010;16:691–8.
44. Simon R (2010) Clinical trials for predictive medicine: New challenges and paradigms. Clin Trials. (In Press).
45. Bonnefoi H, Potti A, Delorenzi M, et al. Validation of gene signatures that predict the response of breast cancer to neoadjuvant chemotherapy: a substudy of the EORTC 10994/BIG 00-01 clinical trial. Lancet Oncol. 2007;8:1071–8.
46. Simon R. Clinical trial designs for evaluating the medical utility of prognostic and predictive biomarkers in oncology. Personalized Medicine. 2010;7:33–47.